NAVAL HEALTH RESEARCH CENTER

PHYSICAL STRENGTH AND PERFORMANCE OF MODERATE DURATION PHYSICAL TASKS

R. R. Vickers, Jr.

Report No. 03-08

Approved for public release; distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (M2)
2300 E ST. NW
WASHINGTON, DC 20372-5300

Physical Strength and Performance of Moderate Duration Physical Tasks

Ross R. Vickers, Jr.

Human Performance Program
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

Report 03-08 was supported by the U.S. Army Medical Research and Materiel Command under Work Unit 60109. The views expressed are those of the author and do not reflect the official policy or the position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution unlimited.

SUMMARY

Problem

Valid models of physical task performance could be useful for selection into Navy jobs and for computer simulations of combat performance. Simple models are desirable, but of little value if they are inaccurate. Previous studies indicated that strength predicted more than 90% of the variance in performance on a wide range of physically demanding U.S. Navy tasks. This finding implies that strength is the only ability to consider for Navy selection and modeling purposes. However, the tasks studied previously lasted at most a few minutes. Work physiology principles predict that longer lasting tasks will have a more complex causal structure in which strength and endurance both are important.

Objectives

The primary goals of this study were to (a) demonstrate that strength is a less powerful performance predictor when tasks last longer than 1 min, and (b) evaluate the hypothesis that the initial strength-performance estimate was biased by the omission of other physical abilities from the predictive equation.

Approach

Structural equation modeling was applied to data from a study of steelworkers. Task performance measures included lifting, carrying, and shoveling tasks lasting 5 min to 15 min. Physical ability measures included the static strength dimension from earlier research and a dynamic strength dimension. Structural equation models were constructed to estimate the relationship between physical abilities and performance.

Results

Static strength strongly predicted performance ($\beta = .86$), but this association was significantly ($p < .001$) lower than the estimate obtained in prior studies of shorter tasks. Adding dynamic strength (i.e., sit-ups, pull-ups) improved overall criterion prediction slightly (semipartial $r = .13$) even though it lowered the estimated effect of static strength by 25% ($\beta = .69$).

Conclusions

The effects of physical abilities on task performance can be estimated accurately only after careful selection of tasks and ability tests. The analysis procedures must provide methods of formulating and testing specific models that combine theoretical considerations with prior empirical findings. Most prior studies of physical abilities and task performance do not meet these criteria. The resulting estimates of the impact of physical ability on performance are likely to be biased. The biases can undermine the accuracy of screening batteries and/or lead to suboptimal performance enhancement interventions.

Introduction

Muscle strength is important for physical task performance. Vickers (1995, 1996) found that muscle strength accounted for 91% of the variance in overall performance for a set of physically demanding U.S. Navy tasks. The virtual 1:1 correspondence between strength and physical performance was surprising because strength is only one of several well-documented physical abilities (Fleishman, 1964; Hogan, 1991; Myers, Gebhardt, Crump, & Fleishman, 1993; Nicks & Fleishman, 1962). These other abilities seem likely to account for more than 9% of the variance in physical task performance. This report examines 2 factors that may have affected the prior estimate.

Task selection may have affected the prior findings. The tasks were chosen because they required strength (Robertson & Trent, 1985). No task lasted as long as 1 min for the average person. Thus, physical capacities such as muscle endurance and aerobic fitness had little opportunity to come into play.

The selection of ability measures may have been important to the prior findings. Indicators were chosen to assess Fleishman’s (1964) static strength dimension. This dimension indicates the maximum force that a muscle group can generate for a brief period. Factor analyses have identified other physical abilities that are logically related to task performance (Fleishman, 1964; Hogan, 1991; Myers et al., 1993; Nicks & Fleishman, 1962). In fact, Fleishman (1964) identified 7 additional factors, including 3 other strength factors (e.g., dynamic strength, explosive strength). A plausible argument could be made that each of these factors could affect task performance. Subsequent work suggests that there may be fewer physical ability dimensions, but consistently indicate more than one. The lack of indicators for these other measures may have biased Vickers’s (1995, 1996) estimates of strength effects. Omissions lead to bias when the missing variables are correlated with both the dependent variable and predictors that are in the model (James, Mulaik, & Brett, 1982). These conditions might reasonably have been met in Vickers’s (1995, 1996) analyses.

This report examines the effects of task selection and ability sampling on ability-performance models. These effects were evaluated by conducting a reanalysis of a study of steelworkers conducted by Arnold, Rauschenberger, Soubel, and Guion (1982). The tasks that are examined here lasted 5 min to 15 min. The strength measures included indicators for Fleishman's (1964) static and dynamic strength dimensions. Arnold et al. (1982) included both dimensions based on a logical analysis of the ability requirements following a job analysis.

This paper reports a reanalysis of Arnold et al.’s (1982) data using structural equation modeling to test explicit theoretical formulations. The analyses focused on two hypotheses. First, the association between static strength and task performance will be $r < .95$. Other abilities (e.g., muscle endurance) should be more important for longer-lasting tasks. As the variance explained by other abilities increases, the proportion of variance explained by static strength must diminish. The second hypothesis was that the estimated relationship between static strength and performance is biased when dynamic strength is omitted from the model. Assuming that dynamic strength would be positively related to both static strength (Nicks & Fleishman, 1962; Myers, Gebhardt, Crump, & Fleishman, 1993) and to performance, the bias should be positive. Adding dynamic strength to the model, therefore, will further weaken the relationship between static strength and performance.

Methods

Sample. Arnold et al. (1982) studied 249 workers (168 men and 81 women) at 3 manufacturing sites in a steel/steel products company. All but 10 participants were steelworkers. The 10 non-steelworkers were included at one research site to increase the variance in ability (Arnold et al., 1982, p. 589). No demographic information other than gender was provided in the original paper.

Sample size for the structural equation models (SEMs) was fixed at $N = 244$. This estimated sample size was the point of convergence for two sets of computations. First, Tables 10, 11, and 12 of Arnold et al. (1982) reported ability-performance correlations for three separate work sites. Each table specified a minimum and a maximum sample size for the reported correlations. A cumulative lower bound for the sample size ($N = 239$) was computed by adding the minimum sample sizes across work sites. A cumulative upper bound ($N = 249$) was computed by adding the maximum sample sizes. The midpoint between the upper and lower bounds was 244. The second set of computations was based on Table 9 of Arnold et al. (1982). This table pooled the strength data across the 3 research sites (Arnold et al., 1982, Table 9). The reported range of sample sizes for the table was 238 to 249. The midpoint of this range (243.5) rounded to $N = 244$.

Simulated Tasks. Different work samples were constructed for each site. The work sample at each site represented tasks that entry-level personnel would perform at that site. Work sample construction for each site was guided by the perception that "…successful task performance in the various positions required general rather than specific physical abilities…"(Arnold et al., 1982, p. 589). Work sample construction also was affected by the view that the "…majority of AWS [abstracted work sample] tasks were related to strength …"(Arnold, et al., 1982, p. 590).

The performance measures for this study consisted of 3 tasks. Each task had been studied at all 3 work sites. Task performance measures were objective, involving a count or weight (e.g., number of bags moved). The other tasks studied by Arnold et al. (1982) either were not studied at all sites or were assessed by ratings rather than direct measurements. Ratings were highly correlated with counts or other direct measures of performance (Arnold et al., 1982, Table 3, p. 591), but restricting the analysis to objectively measured performance on simulated tasks limited attention to those performance measures that were directly comparable to the task simulations investigated in Vickers's (1995, 1996) earlier modeling efforts. The 3 tasks were:

> ***Moving 50-pound (22.7 kg) bags***. Participants lifted and moved bags from one ground-level pallet to another and then back again. Participants were instructed to move as many bags as possible in 5 min. Performance was the number of bags moved.
>
> ***Lifting 75-pound (34 kg) bags***. Participants lifted 75-pound (34 kg) bags to and from a 4-foot (1.2 meter) high table as many times as possible in 5 min. Performance was the number of bags lifted.
>
> ***Shoveling Slag***. Participants shoveled slag into a wheelbarrow until full, then dumped the slag back onto the original pile, and began filling the wheelbarrow again. This sequence was repeated as many times as possible during a 15-min performance period. Performance was the total weight of slag shoveled.

Ability Assessments. Arnold et al. (1982, p. 590) chose ability measures with an emphasis on Fleishman's (1964) static and dynamic strength dimensions. Static strength measures included dynamometer tests for the arm, back, and leg strength. Dynamic strength was assessed by push-ups, leg lifts, pull-ups, and squat thrusts. Push-ups were the total number of push-ups performed "until tired." Leg lifts were the number performed in 30 s. Squat thrusts were the number of thrusts performed in 40 s. Pull-ups were the number performed on a 1.75-inches (4.4-cm) bar "until tired."

Data Extraction. Correlations between measures were taken from different tables in Arnold et al. (1982). Strength measure correlations were obtained from Table 9 (p. 595). That table reported correlations averaged across the 3 sites. Sample size was given as 238 to 249. Ability-performance correlations were estimated by averaging the site-by-site correlations reported in Tables 10, 11, and 12 of Arnold et al. (1982). The midpoint of the sample size given for each table was used in the computations (76, 89, and 79, respectively). Weighted averages of the raw correlations then were computed. This averaging method may slightly underestimate population correlations, but the alternative procedure of using the Fisher r-to-z transformation may introduce a slight bias in the opposite direction (Silver & Dunlap 1987; Strube 1988). The absolute magnitude of the errors and/or differences between the approaches has been modest in simulation studies, so the choice between averaging methods should have little or no impact on the findings. While the r-to-z transformation may be preferable in general, underestimates of the correlations were more acceptable in the present case than overestimates.

Arnold et al. (1982) reported the correlations among task performance measures for just 1 site. Presumably this choice reflected the fact that each work site involved a distinct set of performance measures. The results for a single site probably were reported to conserve space. The original report includes a statement that “Here, as at other sites, these measures [i.e., all of the tasks at that site] were highly intercorrelated” (Arnold et al., 1982, p. 592). It was assumed that the correlations between lifting and shoveling tasks reported in that sample were representative for all 3 samples. The data from that single site, therefore, provided the estimated correlations between performance tasks.

Means and standard deviations for the strength and performance measures were taken from Table 8 of Arnold et al. (1982, p. 594). The values reported for the combined male and female samples were used to correspond to the reported correlations. Those correlations were for the combined samples.

Analyses

SEMs were fitted to the covariance matrix using LISREL 8 (Joreskog & Sorbom, 1993). The factor loading for one indicator variable was fixed at 1.00 in each analysis to establish the scaling for the latent traits. The physical ability dimensions were treated as exogenous variables that influenced an endogenous performance variable.

Model comparisons used several indicators to conform to current recommendations (Boomsma, 2000; Hu & Bentler, 1998; McDonald & Ho (2002). The standardized root mean square residual (SRMR, Joreskog & Sorbom, 1981) and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) were chosen because these indices are sensitive to model misspecification (Fan, Thompson & Wang, 1999; Hu & Bentler, 1998). The nonnormed fit index (NNFI, Bentler & Bonett, 1980; Tucker & Lewis, 1973) was included because it is one of a number of indices that reflect improvements in the fit of the model in a fashion that is analogous to R^2 in regression analyses. These indices are widely reported in the literature. The familiarity of the NNFI and its similarity to R^2 may make this index more readily interpretable than the other indices reported here. Finally, Browne and Cudeck’s (1989) expected cross-validation index (ECVI) was included. This index is a reminder that the results reported in this paper were derived from a single sample. Excessive confidence in the generalizability of models derived in a single sample is a problem in structural equation modeling (MacCallum & Austin, 2000). Both RMSEA and ECVI are relevant to this point because they have population interpretations that are accompanied by estimates of sampling variance. Those estimates can be used to construct confidence intervals as reminders of the uncertainty associated with sampling effects.

Results

Strength Measurement Models. Three strength measurement models based on Fleishman’s (1964) physical ability model were considered. The models were:

A. **Unidimensional (1D)**: All measures loaded on a single ability factor.

B. **Two-Dimensional Orthogonal ($2D_O$)**: Dynamometer strength measures defined one dimension. Leg lifts, push-ups, pull-ups, and squat thrusts defined the second dimension. The two dimensions were orthogonal.

C. **Two-Dimensional Correlated ($2D_C$)**: Dimensions were defined by the same variables as in the $2D_O$ model, but a correlation was added between the two dimensions.

Table 1. Comparison of Strength Models

Model	df	χ^2	SRMR	RMSEA	NNFI	ECVI
Null Model	21	1237.73				
Unidimensional Strength (1D)	14	298.09	.08	.29	.81	1.34
Two-Dimensional Orthogonal ($2D_O$)	14	157.48	.35	.21	.83	.76
Two-Dimensional Correlated ($2D_C$)	13	46.43	.04	.10	.96	.31

Note. See text for details of models. SRMR is the standardized root mean square residual (Joreskog & Sorbom, 1981). RMSEA is the root mean square error of approximation (Steiger & Lind, 1980). NNFI is the Bentler and Bonett (1980) nonnormed fit index also known as the Tucker-Lewis index (TLI; Tucker & Lewis, 1973). ECVI is Browne and Cudeck's (1989) expected cross-validation index.

The $2D_C$ clearly was the best model for each goodness-of-fit index (GFI; Table 1). This model met prevailing standards for acceptable fit (i.e., NNFI ≥ .900, cf., Bentler & Bonett, 1980). This model also met more demanding standards (NNFI ≥ .950) recently recommended by Hu and Bentler (1999). RMSEA fell in the marginal fit range defined by Browne and Cudeck (1989), albeit right at the upper boundary. ECVI criteria have not been definitively established, but it was noteworthy that the lower boundary of the 90% confidence interval for the $2D_C$ ECVI (0.24) was just slightly higher than the ECVI for a saturated model (0.24). Finally, SRMR was less than Hu and Bentler's (1999) recommended cutoff value of .08.

Figure 1 shows the $2D_c$ model. The static-dynamic strength correlation in the $2D_C$ model was r = .76. Standardized residuals suggested several narrow latent traits might be present (arm dynamometer-push-up, z = 3.66; leg dynamometer-back dynamometer, z = 3.26; leg lifts-squats, z = 4.45; leg dynamometer-arm dynamometer, z = -3.11). However, residuals should be viewed with caution until replicated (MacCallum, Roznowski, & Necowitz, 1992). Each of these would be significant applying Green, Thompson and Poirer’s (2001) Bonferroni adjustment procedure.

Performance Measurement Model. The 3 task performance measures defined a single general dimension. Correlations between measures were substantial and approximately equal in magnitude (r = .74 to r = .81). With only 3 strength measures, a unidimensional model fitted the data perfectly. This result was anticipated because 3 indicators are the minimum that uniquely defines a factor. Thus, no other latent trait performance models required consideration. Figure 2 presents the performance measurement model.

Ability-Performance Models. The ability-performance models combined the $2D_C$ ability model with the performance measurement model. Latent trait loadings for indicators were fixed at the values estimated when the measurement models were fitted to the data. This 2-step approach was

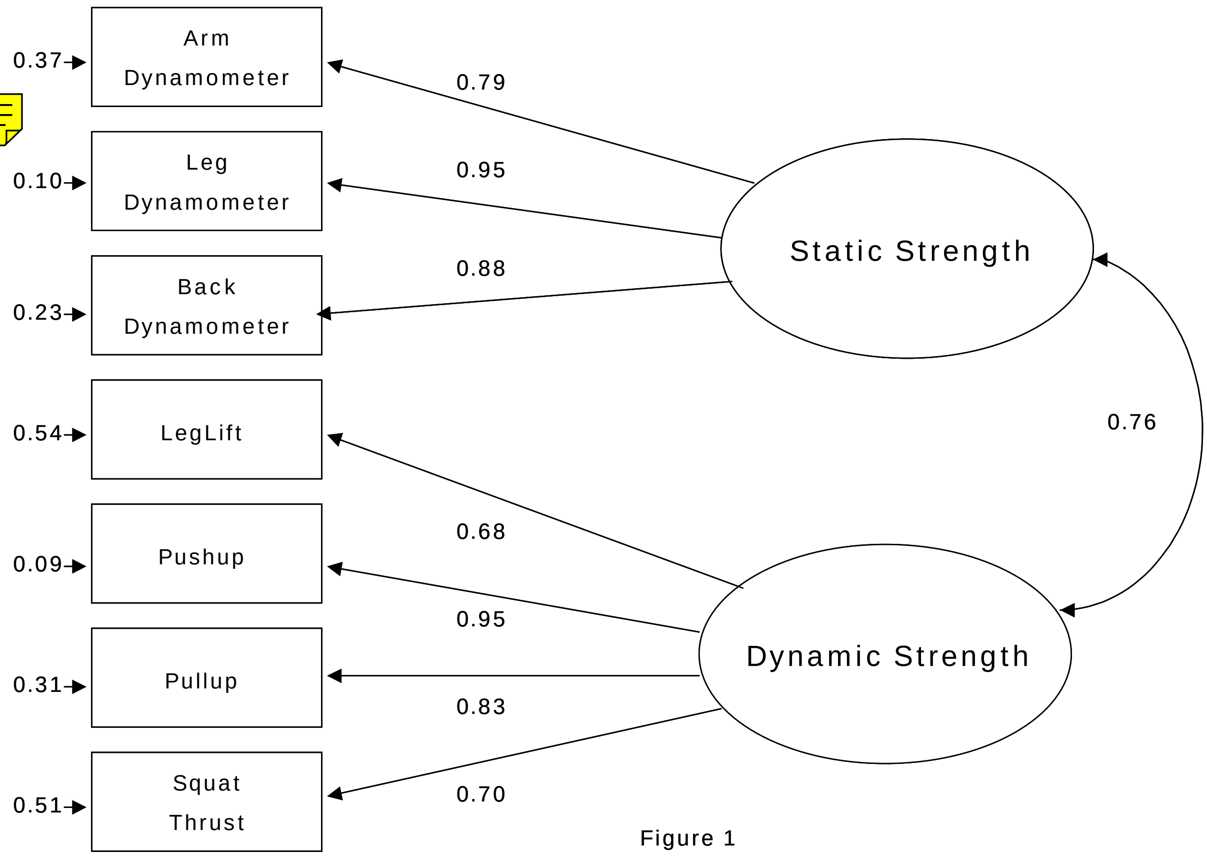

Figure 1

Ability Measurement Model

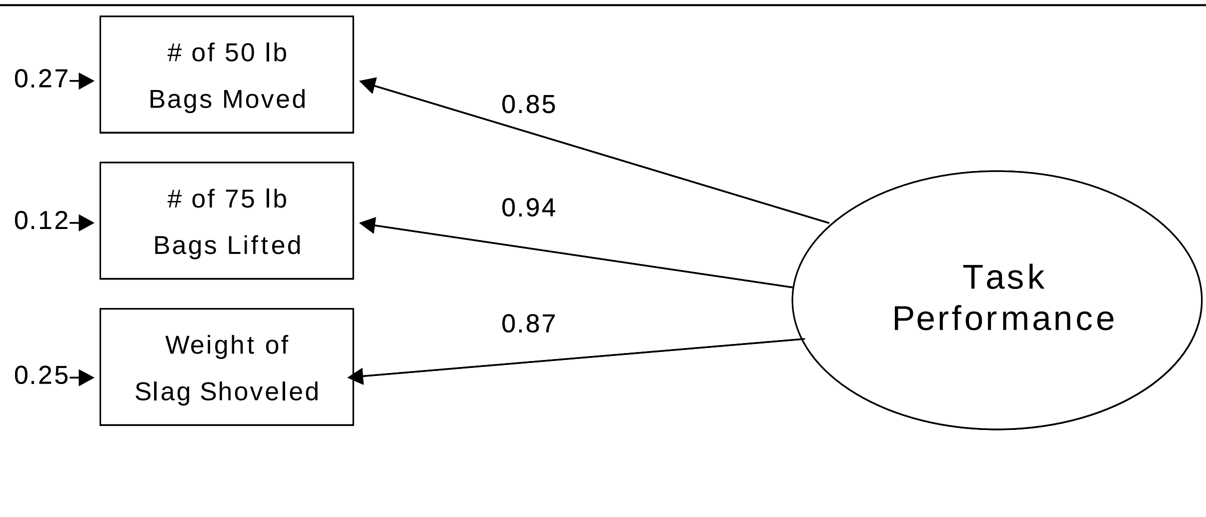

Figure 2

Performance Measurement Model

Table 2. Ability-Performance Models

Model	*df*	χ^2	NNFI	RMSEA	SRMS	ECVI
Null	20	309.32		.12	.37	.99
Dynamic Strength	19	131.99	.59	.088	.092	.65
Static Strength	19	53.38	.88	.061	.047	.43
Static + Dynamic	18	42.43	.91	.052	.036	.38

Note. Model labels indicate the ability-performance relationships in the model. See Table 1 for definitions of column headings.

advocated by Anderson and Gerbing (1988) and recommended by McDonald and Ho (2002). In these analyses, the relationships between the latent traits were the primary concern. The 2-step approach separates the substantive relationships from the specification of the measurement model. McDonald and Ho (2002) showed that this point is important because GFI can be quite different for the two parts of the model.

Ability-performance NNFIs were computed to estimate the fit of the model with attention limited to the ability performance elements of the covariance matrix. NNFI values were computed by subtraction, specifically, χ^2 = 46.43 for the final ability measurement model and χ^2 = 0.00 for the performance measurement model. With the latent trait loadings for the indicator variables fixed at the values estimated when fitting the measurement model, the contribution of the measurement models to the total misfit between the model and the data is equal to the sum of these χ^2 values. The misfit associated with the differences between predicted and observed ability-performance covariances then is computed as χ^2 = (355.75 - 46.43) = 309.32. This computation provided the χ^2 for the null AP model that is reported in Table 2. The NNFI, RMSEA, SRMR, and ECVI values are the actual values recorded directly from the analysis output.

Table 2 presents the substantive models in the order of their goodness of fit to the data. The model with both static strength and dynamic strength fit the data best for each GFI. This model also produced a significant reduction in χ^2 when compared with the next best model (χ^2 = 9.95, 1 *df*, *p* < .001). Figure 3 shows the Static + Dynamic ability-performance model without the details of the measurement models that were previously presented in Figures 1 and 2.

Latent Trait Relationships

Static strength (*r* = .86) was more strongly related to performance than was dynamic strength (*r* = .74). When the 2 strength traits were used as predictors, the standardized regression equation was

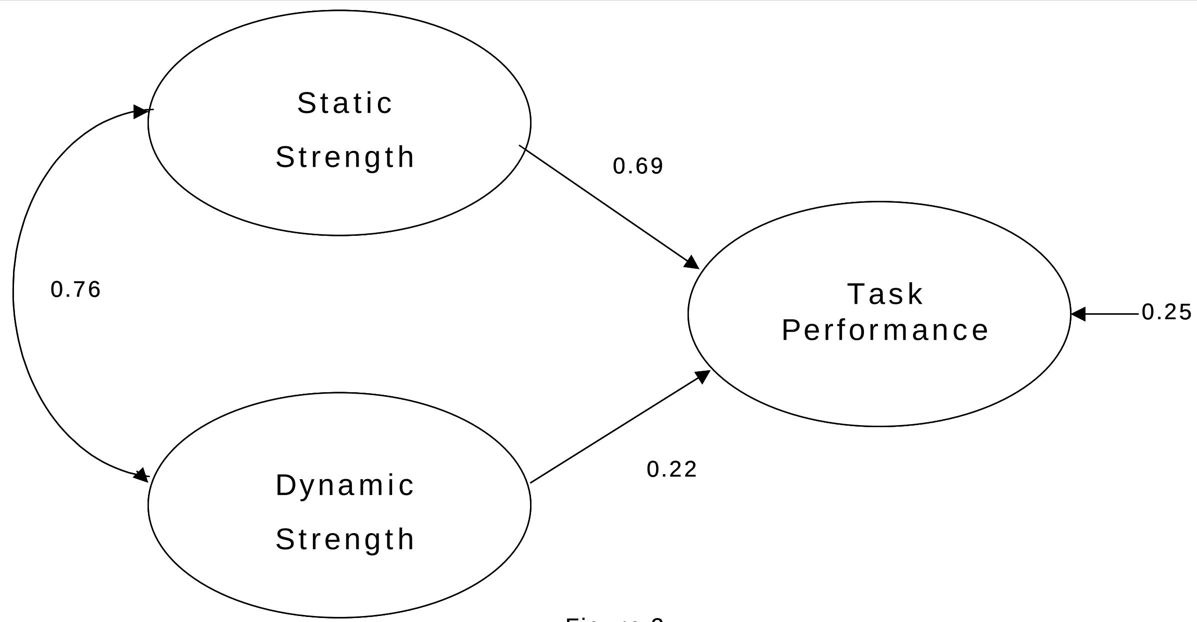

Figure 3

Ability-Performance Model

Performance = (0.69*Static Strength) + (0.22*Dynamic Strength)

Without dynamic strength in the equation, the standardized regression coefficient for static strength was 0.86. The raw regression coefficient for static strength was 1.11 (*SD* = .11) with both predictors in the model and 1.38 (*SD* = .06) with only Static Strength. Thus, adding dynamic strength to the model reduced the Static Strength coefficient by approximately 20% for both the raw and standardized equations.

Correlations between latent traits were examined to estimate the unique amount of performance variance explained by the 2 strength traits. The bivariate latent trait correlations were r = .76 for static strength with dynamic strength, r = .86 for static strength with performance, and r = .74 for dynamic strength with performance. Semi-partial correlations (sr) (Cohen & Cohen, 1983) indicated that the unique contribution of dynamic strength accounted for 1.7% of the performance variance (*sr* = .13) compared with 20.3% for dynamic strength (*sr* = .46). Each correlation exceeded Cohen's (1988) criterion for the minimum effect size that would be of practical or theoretical importance (i.e., $r = .100$).

The residual associations between individual ability tests and individual tasks were examined. The primary objective was to identify any areas of substantial misfit between the model and the data. If substantial misfit was observed, a secondary objective was to determine whether the misfit might be linked to any of the implied latent traits suggested by the correlated residuals in the strength measurement model (see p. 5). An association would be implied if both of the variables contributing to the large residual correlation for the ability measures correlated produced large residuals for one or more tasks. However, all standardized residuals were small ($z = -0.96$ and $z = 0.87$). The model accurately reproduced each of these elements of the overall covariance matrix.

The Task Duration Hypothesis

A direct test of the task duration hypothesis was obtained by comparing the observed correlation between static strength and performance ($r = .86$) to the earlier estimate ($r = .95$; Vickers, 1996). The difference between the correlations was in the predicted direction and highly significant ($z = 8.44$, $p < .001$). Static strength accounted for 16% less performance variance in this study (i.e., 74% vs. 90%).

Discussion

The first study hypothesis was supported. Task duration affects the association between physical ability and physical task performance. The relationship between static strength and performance was significantly ($p < .001$) weaker with tasks lasting 5 min to 15 min ($r = .86$) than with shorter (<1 min) tasks ($r = .95$).

The second study hypothesis was supported. Incomplete sampling of the strength domain biased the estimate of static strength effects on performance. The conditions for bias were met. Static strength was positively related to dynamic strength ($r = .76$). Dynamic strength was positively related to task performance ($r = .74$). Adding dynamic strength improved the ability-performance model, so the inclusion of a dynamic strength effect on performance was reasonable. The standardized regression slope for static strength was 0.86 with dynamic strength omitted from the model and 0.69 with dynamic strength in the model. Thus, omitting dynamic strength inflated the estimate of the static strength effect by 25%.

A correct understanding of the effects of physical ability on task performance requires studies that meet 3 conditions. First, coverage of the ability domain must be broad enough to minimize the risk of omitted variable bias. This condition can be met by ensuring that studies designed to estimate the effect of a specific ability include other ability indicators that are known to correlate with the ability of interest and might reasonably be expected to influence the performance variable of interest (James et al., 1982). Second, the task domain needs to be characterized in more detail. Tasks ordinarily are categorized as lifting, pushing, pulling, carrying, and so on. This type of classification may not be optimal for understanding the relationship between task performance and physical ability. For example, Vickers (1995, 1996) found that the wide range of lifting, pulling, and carrying tasks studied by Robertson and Trent (1985) could be reduced to a single general dimension for modeling purposes. The present findings suggest that task duration may be more important than task type when modeling ability-performance associations in the manual material-handling domain.

Appropriate modeling of the task side of the equation is an overlooked aspect of ability-performance work. Studies of physical tasks have concentrated on tasks chosen because they are critical in some respect. Identifying the most demanding task in a job is an example of how tasks are chosen. Once chosen, each task is treated as a separate criterion. Study findings are task-by-task listings of predictor equations (e.g., Arnold et al., 1982; Robertson & Trent, 1985). This approach has several potential problems. First, there is a greater likelihood that the regression equations will be suboptimal for the population. Chance sampling variation will cause the omission of some useful predictors and/or the inclusion of some irrelevant predictors. These risks are present when a single criterion is considered, but the risk increases with the number of criteria examined. Second, different tasks may require different physical abilities or different levels of the same ability. Any approach that multiplies the number of criteria being considered increases the likelihood of conflict between standards based on different criteria. Third, the criterion-by-criterion approach poses problems if the tasks in a job change. New studies would be needed to set criteria for the new task. This step might not be necessary if the task were seen as one more example of a general performance dimension or a combination of two or more dimensions. A conceptual model of tasks is required for an efficient attack on the problem of setting standards. Vogel, Wright, Patton, Dawson, and Escherback (1980) provided an example of how a conceptual model for tasks can be used in this context, but this approach has not been used as a framework for organizing the empirical evidence. Ultimately, a task categorization might be achieved by sampling tasks systematically to represent different task categories (e.g., lifting, carrying, pulling, pushing) and the workload and duration of the tasks.

The use of appropriate modeling procedures is the third condition that must be met. The formal method of modeling the data also affects the findings. The SEM approach taken here provided formal tests of alternative models. Arnold et al. (1982) employed regression procedures. The present results supported Arnold et al.’s (1982) logical analysis of the ability requirements on the job. This result was obtained using modeling procedures that specifically represented the hypothesized abilities of interest. The results contrast with Arnold et al.'s (1982) conclusion that "…a single measure of arm strength was sufficient for predicting performance on various tasks that call for use of the whole body" (p. 603). Arnold et al. (1982) noted that this finding was incompatible with the complexity of Fleishman's (1964) factor analytic model, but suggested that "…the various types of strength are sufficiently interrelated to allow the identification of a general strength construct" (p. 603). A unidimensional model does not appear reasonable based on the better fit of the two-dimensional model in the ability-performance portion of the covariance matrix. The pattern of residuals indicated that models with more than 2 latent traits were not needed to account for the ability-performance relationships.

Errors in modeling will be more important in some situations than others. Screening job applicants is probably the most common reason for developing ability-performance models. This application does not require a correct understanding of the causal processes involved. The important question is whether the prediction of future performance is accurate. In this case, the omission of a causal variable is important only if it produces a loss of predictive power. The loss depends on how much the addition of the omitted variable would increase the correlation between the predictor composite and the criterion (Rosenthal & Rubin, 1979). This loss depends heavily on the particular variable that is omitted. In the present analyses, omitting dynamic strength would reduce accuracy by 2%.[1] Omitting static strength would reduce accuracy by 20%. The first loss could be disregarded in many cases; the second loss would be hard to ignore. A complete ability-performance model can be useful in these cases as a guide to ensure adequate coverage of the ability domain relative to the tasks of interest in a given situation.

Modeling errors are more critical when formulating interventions. In these cases, a correct causal[2] model is needed for an accurate forecast of the effect of the intervention. For example, consider a hypothetical program designed to increase dynamic strength. Such a program might be a standard physical conditioning program involving push-ups, squat thrusts, and so forth. The expected payoff from this program would be substantial if the expectation was based on the bivariate relationship between dynamic strength and task performance observed in this study (i.e., $r = .74$). However, the complete model suggests that the actual effect will be less than one third of this expectation (i.e., $\beta = 0.22$) if the program truly affected only dynamic strength. Both types of strength would have to be measured to

[1]The gain would be substantially larger if static strength were added to an existing battery of dynamic strength measures. The bivariate relationship between dynamic strength and performance was much smaller than that between dynamic strength and performance.
[2]Causal interpretations of covariances must be viewed cautiously. Regression coefficients cannot be interpreted routinely as indicators of the magnitude of causal effects (Sobel, 1996). In this instance, strength probably meets any reasonable criteria for a causal influence on performance.

determine the actual effect of the program. This information could be critical in refining the program to isolate and accentuate the "active ingredient" that produces performance effects. Thus, a sound statistical model is important for the development and evaluation of intervention programs.

This study employed combined data from males and females. Combining the sexes increased the sample variability in physical ability tests and performance relative to the within-sex variation (cf., Arnold et al., 1982). The observed correlations, therefore, will be larger than they would be in a single-sex sample (Hunter & Schmidt, 1990). If this tendency extends to the estimated correlations between latent traits, the present study underestimates the shrinking effect of task duration on the static strength-performance relationship. Vickers's (1995, 1996) estimate of the relationship for short duration tasks was based on analyses that separated males and females. However, the effects of range differences might be absorbed in the measurement model without affecting the ability-performance relationship. This issue needs further study.

This study extended prior evidence (Vickers, 1995, 1996) that strength is a strong predictor of performance on physically demanding occupational tasks. However, strength must be considered in the context of a full representation of physical abilities to reduce the risk of obtaining biased estimates of strength effects. A detailed investigation is worthwhile even though the correlations between strength dimensions are moderate to strong. The results also were consistent with the view that static strength is less important as task duration increases. This result is common sense and consistent with muscle fatigue research showing that stronger individuals fatigue more rapidly than weaker individuals (e.g., Clarke, 1986). The implication of this common sense observation is that a systematic understanding of the task domain is critical for understanding ability-performance relationships. Further work is needed to determine whether tasks can be represented as multidimensional variables with different performance dimensions that correspond to different elements of ability models. A one-to-one mapping of task characteristics onto physical abilities would lead to models such as Vogel et al.’s (1980) translation of U.S. Army tasks into strength and aerobic fitness requirements. However, this type of framework may be less effective than one that treats task performance as a distinct domain. Tasks have attributes such as the method of performance and effects of experience that may need to be represented in models to fully understand ability-performance relationships. The primary result of this study, therefore, is that it indicates the need for systematic exploration of both sides of the ability-performance equation to optimize selection and intervention practices.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology, 67*, 588-604.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606.

Boomsma, A. (2000). Reporting analysis of covariance structures. *Structural Equation Modeling*, *7*(3), 461-483.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*, 445-455.

Clarke, D. H. (1986). Sex differences in strength and fatigability. *Research Quarterly for Exercise and Sport*, *57*(2), 144-149.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*(1), 56-83.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice-Hall.

Green, S. B., Thompson, M. S., & Poirer, J. (2001). An adjusted Bonferroni method for elimination of parameters in specification addition searches. *Structural Equation Modeling*, *8*(1), 18-39.

Hogan, J. C. (1991). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76*, 495-507.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park: Sage Publications.

James, L. R., Mulaik, S. M., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201-226.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The dimensions of human physical performance: Factor analysis of strength, stamina, flexibility, and body composition measures. *Human Performance, 6*, 309-344.

Nicks, D. C., & Fleishman, E. A. (1962). What do physical fitness tests measure? A review of factor analytic studies. *Educational and Psychological Measurement*, *22*(1), 77-95.

Robertson, D. W., & Trent, T. T. (1985). *Documentation of muscularly demanding job tasks and validation of an occupational strength test battery (STB)* (NPRDC Tech. Rep. No. 86-1). San Diego, CA: Navy Personnel Research and Development Center.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*(5), 395-396.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*(1), 146-148.

Strube, M. J. (1988). Averaging correlation coefficients: Influence of heterogeneity and set size. *Journal of Applied Psychology, 73*(3), 559-568.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research*, *24*(3), 353-379.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Vickers, R. R., Jr. (1995). *Physical task performance: Complexity of the ability-performance interface* (NHRC Tech. Rep. No. 95-30). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (1996). *Generalizability test of a physical ability-job performance model* (NHRC Tech. Rep. No. 96-16). San Diego, CA: Naval Health Research Center.

Vogel, J. A., Wright, J. E., Patton, J. P., Dawson, J., & Escherback, M. P. (1980). *A system for establishing occupationally-related gender-free physical fitness standards* (USARIEM Tech. Rep. No. 5). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY) 26 Sep 2002

2. Report Type Interim

3. DATES COVERED (from - to) Sep 2000 - Aug 2002

4. TITLE AND SUBTITLE
Physical Ability and Performance of Moderate Duration Physical Tasks

5a. Contract Number: US Army Reimb
5b. Grant Number:
5c. Program Element:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60109

6. AUTHORS
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Chief, Bureau of Medicine and Surgery
MED-02
2600 2300 E St NW
Washington DC 20372-5300

9. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 03-08

10. Sponsor/Monitor's Acronyms(s)

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

Previous research has demonstrated a strong relationship (r = .953) between strength and performance on physically demanding U.S. Navy tasks. The earlier work covered tasks lasting <1 min and limited strength measures to static strength. The relationship might be weaker for longer-lasting tasks that required muscle endurance. The earlier estimate also may have been biased by the omission of other aspects of muscle strength. This study related static and dynamic strength to performance on physical tasks lasting 5 to 15 min to test these hypotheses. As expected, the association between static strength and task performance was significantly (p < .001) weaker than in the earlier study (r = .86). Omitted variable bias was indicated by the fact that the regression coefficient relating static strength to performance shrank by 25% (i.e., β = .69) when dynamic strength was added to the model. Sound models require systematic sampling on both sides of the ability-performance equation. Incompleteness is especially important when a model is the basis for performance enhancement interventions. In this specific instance, the results suggest that omitted variable bias would lead to a 25% overestimation of program effects.

15. SUBJECT TERMS
physical fitness, aerobic fitness, run tests, meta-analysis, validity

16. SECURITY CLASSIFICATION OF: a. REPORT	b.ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UNCL	UNCL	UNCL	UNCL	17	Commanding Officer
					19b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429